DOCUMENT RESUME

ED 463 293 TM 033 715

AUTHOR	Kadamus, James A.
TITLE	Standards to Assessments: Looking at the Whole Picture. Memorandum.
INSTITUTION	New York State Education Dept., Albany. Office of Elementary, Middle, Secondary, and Continuing Education.; New York State Library, Albany.
PUB DATE	2001-09-00
NOTE	27p.; For 1998 Memorandum, see TM 033 713. For 1999 Memorandum, see ED 446 121.
AVAILABLE FROM	For full text: http://northport./K12.ny.us~nms/standards.pdf.
PUB TYPE	Guides - Non-Classroom (055)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Decision Making; Educational Testing; Elementary Secondary Education; *Reliability; *Scores; State Programs; State Standards; *Test Construction; Test Use; Testing Programs; *Validity
IDENTIFIERS	*New York

ABSTRACT

This field memo is the third in a series of updates from the Deputy Commissioner of the Office of Elementary, Middle, Secondary and Continuing Education to all teachers and administrators of public and nonpublic schools on standards and assessments in New York. This document explains how the New York State Education Department determines if test scores are fair, reliable, and valid, and how cut scores are set. It also discusses how test data can be used to inform decision making at various levels in the school community and contains a section of commonly asked questions with answers. The discussion is set in the context of the new state assessments developed in accord with New York's standards of learning. The test construction process is described from the development of test items by New York educators through the statistical checks that assure reliability, validity, and fairness. The booklet also discusses using the data to inform instruction and develop instructional policy. It suggests ways parents can use the data in working with their children's teachers. Attachments contain the draft schedule for assessments in 2002, a chart of the test development process, and a list of approved alternative assessments. A glossary is attached. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 463 293

MEMORANDUM

New York State Department of Education
The University of the State of New York

Standards to Assessments: Looking at the Whole Picture

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

L W. Stanton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033715

Deputy Commissioner
Office for Elementary, Middle, Secondary and Continuing Education
Albany, NY 12234

BEST COPY AVAILABLE

THE STATE EDUCATION DEPARTMENT/THE UNIVERSITY OF THE STATE OF NEW YORK/ALBANY, NY 12234

James A. Kadamus, Deputy Commissioner
Office for Elementary, Middle, Secondary and Continuing Education
Room 875 EBA (518) 474-5915

September 2001

TO: All Teachers and Administrators of Public and Nonpublic Schools

FROM: James A. Kadamus

SUBJECT: Standards to Assessments — Looking at the Whole Picture

This field memo is the third in a series of updates on the standards and assessments. It explains how the Department determines if test scores are fair, reliable, and valid and how cut scores are set. It also discusses how test data can be used to inform decision-making at various levels in the school community, and contains a section on commonly asked questions and provides the answers. There are also a few attachments at the back of this field memo that provide more information related to assessments.

Many of you have shared with me and my colleagues in the Department your thoughts and understandings of these highly technical processes. Without a doubt, the language can be confusing, and it is not uncommon for a term to mean one thing to one person and something different to another. In an effort to clarify terminology, a glossary of test development- and assessment-related terms is included. We think it is important for every teacher and administrator to have common definitions for these terms and to understand the processes of test development and scoring. It will be useful as you read the Connecting the Scores to the Standards and Setting Cut Scores sections of this field memo.

The results on State tests indicate continued achievement of the standards. More students are taking the tests and demonstrating increased proficiency in reaching the standards. We have had widespread improvement in Regents results, a confirmation that the standards are guiding classroom instruction. In many cases, students are taking and passing Regents examinations before they are required to do so. This is all very good news. The test data are also telling us where the gaps are in student performance. This will enable educators in all parts of our system to direct resources to revise curriculum and instruction so that all students can meet the learning standards.

The score reports on the new State assessments and School Report Cards provide a wealth of information on student performance and school and school district performance. There is much data to be digested and to be used for decision-making purposes by a variety of stakeholders (i.e., school boards, administrators, and teachers). We have provided some examples of how the data can inform decision making.

Throughout this field memo, we also discuss the standards, curriculum development, and classroom instruction and how these components connect and lead to the assessments. We view the standards as the organizing principles of the State assessments, and the State assessments as the accountability mechanism of the standards. In between the two, curriculum development and classroom instruction, aligned to the standards, bring the standards to life and provide the mechanisms for student learning. These four elements, the standards, curriculum development, classroom instruction, and assessment, are essential components of our education reform effort.

We have come a long way since 1992-93, when the first draft of the State learning standards was sent to the schools. We have passed the half-way point in the nine-year phase in of the new standards and assessments. Teachers are working with their colleagues to develop standards-based curriculum and lesson plans that require students to use critical-thinking and problem-solving skills in real-life situations. The revised State assessments are challenging students to demonstrate their achievement of the standards. What began as draft documents detailing what students should know and be able to do, has evolved into a seamless integration of standards, curricula, instruction and assessments.

I trust that this update will help us all to keep the whole picture in mind as we continue to focus our efforts on greater student achievement.

Attachments

A: The New York State Testing Program:
Current Status of Assessments
B: Test Development Process
C: Approved Alternative Assessments
D: Glossary

Connecting Scores to the Standards:

How do we ensure that each State test is a valid and reliable measure of the achievement of the learning standards and is fair to all students?

The development of a State examination is a lengthy process requiring many types of expertise. In the initial test development phase, New York teachers study the State syllabi/core curriculum guides and learning standards and develop specifications, or a "blueprint," for each test. Next, committees of New York teachers and content area specialists draft and refine test questions and scoring rubrics which fit the test blueprint. After a sufficient number of questions have been written, they are "pretested" with small representative samples of the State's students. Results from these pretests are statistically analyzed to determine question difficulty, fairness, and appropriateness for inclusion in a test. Next, "field tests" are developed, using pretested questions, and administered to large representative samples of students across the State. Statistical analysis of the field test results ensures that different test forms are comparable in difficulty, reliable, and appropriate in length. Following the field-testing, final test forms are assembled and fine-tuned using information on testing characteristics determined from the field tests.

We are confident in the validity of the State assessments. For each examination, there is a body of statistical evidence that confirms that the examination measures one important skill. That skill, by design, is inferred to be the particular learning standards that the examination addresses. The recent construct study -- a study of the characteristics that a test is designed to measure -- shows that the Regents examinations in Mathematics A and Comprehensive English faithfully measure an array of cognitive skills from the standards in those subject areas.

Another element of ensuring fairness centers on *how* the test measures what it measures. In order to be fair, a test should be free from questions or concepts that are culturally inaccessible to part of a population, offensive, stereotypical, or demeaning. In order to ensure this, New York State test questions undergo sensitivity reviews by certified experts trained on New York State Sensitivity Guidelines as well as bias reviews by content experts who are informed by specific statistical analysis to uncover unwanted/unexpected differences between population subgroups. Each question is reviewed for its portrayal of human dignity as well as for statistical bias. Questions that are found to have biases or sensitivity concerns are revised to delete unwanted effects or dropped altogether if satisfactory revision is not feasible.

There are several ways to regard reliability, but all have in common an evaluation of the relationship between the true skill or ability of the child and what the test says about that skill or ability. One source of reliability, inter-rater agreement, is very high for all of the State examinations. We take this to be the agreement of scores given through application of rubrics across scorers, either agreeing exactly or agreeing within one point. Normally, this agreement on State examinations is in the mid .9 range.

BEST COPY AVAILABLE

A second major source of reliability is the internal consistency of the examinations. State examinations with essay components typically have reliability rates of .89 to .94, while tests having larger multiple-choice components have even higher reliability ratings.

How are passing scores for tests such as the Regents Comprehensive Examination in English determined?

Setting Cut Scores:

It is generally agreed that "passing" ought to be what a "proficient" student should know or be able to do in a subject. However, determining where this point is on a particular test is much more difficult than counting up the number of questions answered correctly. To determine the passing score, a formal standard setting study is conducted based on the reasoned judgment of subject matter experts and student performance. The performance of students on the tests is at first measured on a logarithmic scale. This is very technical. Student scores are placed on the same scale by placing students according to their chances of answering each question correctly. For the Regents examinations, this must be converted onto a scale that the Board of Regents has defined, in which 65 is passing and 85 is passing with distinction. To perform this conversion, a standard setting study is conducted. The expert panel of teachers first defines what it means to pass and what it means to pass with distinction. They then determine which test questions must be passed to meet those definitions. Because student performance is also on the scale, the points at which these items divide correspond to points of student performance. The two points of passing and passing with distinction are then algebraically mapped on to a scale. For the Regents examination, this scale ranges from 0-100 with 65 representing the passing point and 85 representing the passing with distinction point.

The State Education Department goes to great lengths to ensure the trustworthiness of scores students receive on State assessments. It is of critical importance that all interested parties — board members, administrators, teachers, students, parents, community leaders, and higher education administrators — have confidence that the test questions are fair, valid, reliable, and free from bias, and that the scores students receive accurately reflect their achievement of the learning standards.

Although a State test score is one important indicator of achievement, it is the Department's recommendation that schools provide multiple measures to assess student achievement of the learning standards. The State assessments provide a single uniform measure across all school districts and classrooms. That information is most fair and reliable when large numbers of students are tested on large numbers of items. To understand an individual student's needs, the State test should be used in conjunction with other appropriate measures.

BEST COPY AVAILABLE

Using the Data:

How can schools/school districts use the data to inform instruction and develop instructional policy that will increase student achievement of the learning standards?

Testing is a data-rich activity. We look to the data to tell us how many students took a test, how many were successful, how many were students with disabilities or English language learners, and how many in a given cohort must still take the test. And the list goes on; these are just a few of the ways in which we look at student performance data.

As noted above, student performance data, or test scores, provide a wealth of information, not only about individual students and cohorts of students, but also about school building performance, school district performance, regional results, and statewide results. Score reports and School Report Cards are valuable tools for putting together classroom, building, district and statewide pictures of student achievement.

Standards drive changes in curriculum and classroom instruction across the State. Student performance data can inform decision making about school district expenditures, instructional programs, academic intervention, professional development, and, in some cases, school staffing. These decisions can ultimately help students meet the State learning standards.

Rather than looking at student performance data only in terms of student achievement, we can also look at this data as a powerful management tool. From that perspective, performance data provide knowledge on which to base educational policy and make effective decisions in support of the standards.

Data-Informed Decision Making

Student performance data indicating how well your district, school, classroom, or student is achieving the standards should inform instructional programs. It can be a management tool to influence educational decisions on macro and micro levels. The following examples are ways in which educators can use the data to make effective decisions to help students meet the standards.

In the **classroom** (micro level), an individual student score report provides information on the student's performance. Because the report provides feedback on how the student performed overall as well as on the specific parts of the test, it can be used as one measure to tailor academic intervention services to meet the specific instructional needs of the student. Collectively, the individual student score reports illustrate the level of success of the teaching strategies used in the classroom. Additionally, teachers may get item analyses by individual student and student group from their Regional Information Center located at the BOCES or from their central office information center in the Big 5 school districts. Teachers can use these reports to analyze the effects of their programs and make adjustments to serve their students better and to identify students who are meeting or exceeding the standards. They may seek out professional development to strengthen their standards-based teaching. In summary:

BEST COPY AVAILABLE

- ***Teachers*** can use student performance data as a means of:

  - ✓ Evaluating their lesson plans/curriculum for areas of misalignment with the standards.
  - ✓ Revising their lesson plans/curriculum to ensure alignment with the standards.
  - ✓ Identifying students who are at risk of not achieving the standards.
  - ✓ Contributing to the development of a plan for academic intervention services for each student in need of such services.
  - ✓ Recognizing their need to engage in professional development activities designed to support the learning standards and strengthen their classroom instructional strategies.
  - ✓ Identifying and recognizing students who are meeting or exceeding the standards.

In the **district and its buildings** (macro levels), school and district summary reports provide a broader picture of student achievement on a given assessment and define where attention or more resources are needed for program or classroom instruction. This information can guide decision makers in determining if staff need professional development to fully teach to the standards; if the curriculum needs revision; or if there are major systemic problems that require a curriculum overhaul. It can also be a means of identifying if other programmatic attention is needed (i.e.,additional classes and personnel to provide academic intervention services; increased resources such as textbooks, microscopes, computers, or library resources; or physical plant issues such as heating, cooling, or safety/security-related problems that can negatively affect student learning). On a positive note, this information can identify teachers whose outstanding classroom instruction has helped students meet or exceed the standards. Recognition of teacher and student achievement can then be celebrated. In summary:

- ***School Administrators*** can use student performance data to make decisions that relate to:

  - ✓ Identifying areas of strength in the standards within their programs that can be built on as well as areas that need to be strengthened.
  - ✓ Expanding or adding classes to accommodate academic intervention services and support services.
  - ✓ Hiring personnel to staff additional classes or to provide more teaching assistants for expanded classes.
  - ✓ Updating technology and other equipment such as microscopes, etc.
  - ✓ Updating library resources.

- ✓ Providing professional development opportunities for teachers, including opportunities for teachers to share best practice, as well as support for summer retraining and skill enhancement for teachers whose students have low achievement levels.
- ✓ Supporting mentor programs for new teachers.
- ✓ Celebrating success and recognizing achievement of teachers and students.

At home, the individual student score reports for the elementary and intermediate English language arts and mathematics tests provide parents and guardians with feedback on how their child performed overall and on each part of the tests. These reports can be useful because they pinpoint areas of strength and identify where additional instruction is warranted. Reports of a technical nature are sometimes confusing. Parents should always contact their child's teacher or school administrator if they need assistance interpreting the score report. For example, if a child scores a high Level 2 on the elementary English language arts test, it might mean that the child is struggling in only one or two areas and requires limited instructional intervention. However, if a child scores a low Level 2 or a Level 1, he/she needs serious additional instruction in those areas that present the greatest challenge to achieving Level 3, which is the score level that indicates meeting the standard. Parents and guardians should celebrate the success of students who achieve Levels 3 and 4 with appropriate rewards. In summary:

- ➢ ***Parents*** can use the performance data for their child as a means of:

  - ✓ Conversing with their child's teacher(s) to ensure continuous progress in meeting the standards.
  - ✓ Working with their child's teacher to ensure that he/she receives the academic intervention or support services needed to meet the standards.
  - ✓ Working with their child to make sure he/she gets practice at home on areas of the standards.
  - ✓ Helping their child, with guidance from the teacher, to focus his/her attention on those standards that are the most challenging to achieve.
  - ✓ Praising and rewarding performance that meets or exceeds the standards.

School board members and central administration staff will find the score reports and School Report Card useful as they make decisions about school budgets, staffing, acquisition of resources, and facilities planning. Districts that are in high need areas and whose students are at risk of not meeting the learning standards have the greatest need and the farthest to go to meet the learning standards. Making such great strides may mean reallocating resources to areas of greatest need and making hard decisions about instructional programs, staffing, and professional development. In high need areas where schools have met with success and student performance is high, school boards and administrators will want to recognize and showcase that success. In summary:

- **School board members and central administration** can use student performance data to focus district-wide goals and objectives, such as :

  - ✓ Making long-range plans to support continued student achievement.
  - ✓ Interpreting student performance trends year-to-year to support program needs.
  - ✓ Planning programs for students in need of particular help in reaching the standards.
  - ✓ Identifying school/district programs that need strengthening.
  - ✓ Gaining support for reallocation of district funds.
  - ✓ Celebrating success/recognizing achievement.

District score reports and School Report Cards illustrate for **local businesses and the community** the effectiveness of the instructional programs in their local school district. Businesses have a vested interest in seeing their communities prosper. Many benefits can derive from partnering with their local school districts to provide representation on various committees (i.e., parent-teacher-student association, shared decision-making committee, board of education, and Committee on Special Education). Their business perspective is valuable as schools seek to identify creative ways of stretching resources or providing nontraditional avenues for programmatic and instructional support. This information can be a catalyst for supporting increased student achievement and for recognizing outstanding achievement by students and the district in meeting the learning standards. Businesses can make contributions to the school district to help purchase needed instructional resources, hire additional staff, support student activities to recognize outstanding performance, create work-based learning opportunities, or offer scholarships for college study or job opportunities with flexible work schedules for students who must work and still need to meet their schedule for academic intervention services. In summary:

- ***Business/Community*** can use student performance data as a means of:

  - ✓ Partnering with schools to identify areas in which the business/community can support or recognize student achievement to meet the standards.
  - ✓ Supporting adequate funding for local schools to provide needed staff and instructional resources.
  - ✓ Conveying to the schools their expectation for the knowledge and skills that future employees should bring to the job.
  - ✓ Understanding the level of knowledge and skills future employees will have.
  - ✓ Creating work-based learning opportunities that reinforce and apply knowledge and skills learned in school.
  - ✓ Providing scholarships for outstanding students to support continued study.
  - ✓ Providing flexible schedules for working secondary students to accommodate instructional needs.

As we look to higher education to provide teachers who have been educated using standards-based instructional materials and can incorporate the standards into their teaching strategies, **higher education administrators** are looking to the schools to provide students who demonstrate high academic performance by meeting the learning standards. It is, in some ways, a symbiotic relationship focused on the standards and high achievement. Student performance data illuminates the successes and needs of new teachers as well as students. In summary:

- ***Higher Education Administrators*** can use student performance data as a means of:

  - ✓ Selecting candidates for admission.
  - ✓ Placing entering students in appropriate programs.
  - ✓ Ensuring that teacher education curricula includes instruction in strategies for using the State learning standards for curricular and lesson plan development.

Answering Your Questions About the State Testing Program:

Many of our colleagues have called, emailed, or written to us with questions related to the State assessments. Some have even handed us lists of questions at regional forums. Below are the most frequently asked questions and the answers.

Q 1: ***How are the schedules for the administration of State assessments established?***

A: School districts and professional organizations are consulted regularly. Their responses are compiled and discussed with advisory groups to help determine schedules that best accommodate all students.

Q 2: ***What is the testing schedule for 2002?***

A: The testing schedule is listed on Attachment A.

Q. 3: ***How do the revised Regents examinations compare in difficulty to Regents examinations of the past?***

A: As was true of Regents examinations in the past, a passing score (65 or higher) on the revised Regents examinations represents a level of knowledge and skill in each subject that prepares the student to begin college-level study or skilled employment. Based on the learning standards, the revised Regents examinations require more conceptual understanding, problem-solving, application of knowledge, and critical analysis than previously required.

It is important to keep in mind that the low-pass score (55-64) in place during the phase-in period represents a level of achievement significantly higher than that of the Regents Competency Tests but not comparable to the traditional Regents level of performance.

Q. 4: ***How does the scoring of the revised Regents examinations compare to the scoring of the Regents examinations of the past? How can a raw score of less than half of the total points on the test result in a passing score?***

A: The passing score on each revised Regents examination is established with reference to the learning standards in that subject. The passing score of 65 is set by a formal standard-setting study of performance needed to demonstrate achievement of the learning standards. Each form of the test is equated so that the same scale score represents the same level of achievement. Because the test forms vary somewhat in the mix of easier and more difficult items, the relationship of the raw score to the scale score also varies. That is why conversion grids are provided with each test for translating raw scores to scale scores.

Q 5: What is the Department's guidance to school districts in choosing instruments to test students who performed poorly on State examinations?

A: Districts must first be careful to choose instruments that directly assess the learning standards. When selecting instruments for use in grades in which the State assessments are not given, districts should also account for their own particular sequencing of the instructional program. The most appropriate instrument for one district may not be appropriate for another. Studies of the suitability of instruments must be done in a formal, deliberative manner, using the best content experts available.

Q 6: What are the overriding considerations in accommodations for students with disabilities?

A: It is essential that testing accommodations support the validity of the test scores of students with disabilities in that they preserve the constructs being measured while, at the same time, permit students with disabilities to participate in testing programs on an equal basis with their nondisabled peers. Test accommodations should be based on the individual needs of the student and should not be automatically provided or restricted for all students with a particular disability or in the same program/placement. For example, time should not automatically be doubled because the student has a learning disability. Only students whose Individualized Education Program (IEP) or Section 504 Accommodation Plan includes test accommodations or students declassified by the Committee on Special Education, who have on their last IEP that testing accommodations are to continue, are eligible for testing accommodations. Only those accommodations specified on the student's Individualized Education Program (IEP) or Section 504 Accommodation Plan may be provided to a student taking a State examination.

Q 7: How long does it take to develop a State test?

A: It generally takes about two years to develop a new State test. The process is outlined in Attachment B.

Q 8: Will State assessments continue to be paper and pencil?

A: Yes, for the time being. The State does continue to try out different formats for tests. For example, a representative cross-section of schools was selected to test out the feasibility of web-based assessment using the Intermediate Technology Assessment Sampler.

Q 9: How does the State determine the amount of time allotted to complete an examination?

A: One of the purposes of pretesting and field-testing is to see how much time it actually takes the students to complete the responses. The testing time determination is based on this information.

Q 10: Does the time limit affect the student's performance on the test?

A: No. The Department conducted an evaluation of data from the January/February 2000 elementary English language arts test to determine whether the factors of the time allotted for the test (speededness) and fatigue had an impact on student performance.

Department staff also evaluated whether the placement of certain test questions on specific parts of the three-day test contributed to factors of speededness and student fatigue.

The evaluations found that neither the time limit nor the placement of certain questions influenced the scores.

Q 11: What is a sensitivity review and how is it used in test development?

A: The goal of sensitivity review is to ensure that test materials do not portray individuals or groups in any unfair or negative way. The model for sensitivity review used by the Office of State Assessment has four stages:

1) develop guidelines for use in reviewing all test material;

2) teach educators, parents, and other appropriate professionals how to apply these guidelines reliably;

3) review all test material, instructions, questions, scoring rubrics, interpretive booklets, etc., for adherence to the guidelines; and

4) approve or reject test questions proposed for use in New York State examinations on the basis of the guidelines.

Questions that do not meet sensitivity guidelines are not used on State assessments.

Q 12: What is bias analysis and how are State assessments screened for bias?

A: Bias analysis evaluates whether a test question is asking the same thing with the same degree of difficulty for one group of examinees as it is for another group with the same level of skills. Bias analysis has two components: an empirical evaluation of the characteristics of test questions and a judgment concerning the results of that evaluation.

The Office of State Assessment uses the Mantel-Haenszel procedure to evaluate test questions and the derivatives of this procedure are used for essay questions. This procedure hypothesizes that examinees of the same overall skill level should have the same probability of answering correctly all questions that measure that skill.

Statistics derived from this process can show how a question that is very difficult for one group of examinees should be very difficult for another group, and one that is very easy for one group should be very easy for all. When a question is difficult for one group of test takers and easy for another group that has the same level of skills, then the question may be measuring some characteristic of a particular group of examinees that is unrelated to its intended purpose.

Bias analyses are conducted on New York State examinations after the field-testing. Items that are found to be biased are not used on a State examination. In addition, once the statistical analyses are completed, content specialists who review the results can begin to see patterns in the data. This will inform future item and test development.

Q 13: What is the relationship of readability to item difficulty?

A: Readability is one method of anticipating or predicting the difficulty of a reading passage for a particular group of students. In the New York State Testing Program, readability is one criterion used to select the appropriate passages for pre-testing. However, the decision on which passages fit in the desired range of questions for a test is based on the pretest and field test performance that indicates how accessible each passage was to students.

Q 14: How is reliability determined on the Regents examinations and what levels of reliability are those examinations meeting?

A: 1) As stated earlier in this memo, overall reliability focuses on the consistency of a student's performance across measures and time. Naturally, the greater the breadth of the measure, i.e., the more test questions, the greater the reliability. Reliability is best achieved by evaluating performance based primarily on the whole test before considering smaller portions of the test.

2) Inter-rater reliability is essential for Regents examination scoring and there are two methods used by the Department for analyzing inter-rater reliability:

- The most important means for ensuring reliability of scores on the Regents examinations is by using uniform training and scoring procedures. Scorers all receive training on the same materials. Regents examinations are team scored by teachers who have received the scoring training. The assigned score for each essay is the average of two scorers with a third scorer involved if the two scores are more than a point apart.

- Additionally, a 10 percent audit of Regents examinations is conducted by the Department after the test administration period.

Q 15: Why are the results of the reliability and validity studies on the new State assessments considered good and defensible ?

A: The validity and reliability studies are conducted repeatedly with each test administration and are, therefore, replicable. Because they are replicable, they are defensible. The studies are reviewed by expert committees for their soundness. These committees include both the Technical Advisory Group, composed of top psychometric experts in the nation, content experts from around the country, and ad hoc blue ribbon committees appointed by the Commissioner for particular studies.

The major reason why the quality of the examinations is upheld in research is that the process of development, including repeated analysis and review by experts of wide-scale trial testing, eliminates poor test questions or poor stimulus materials (e.g., reading passages) before they can affect student results. Further, the State is vigilant in reviewing all comments about the examinations from teachers and other educators and studies each comment carefully to catch any errors and amend them. The major safeguard in that function is the release of the examinations after administration so that they can be held to public scrutiny.

Q 16: Why is teaching to the test not considered an effective use of instructional time?

A: State tests use new items with each administration, except in the case of a few restricted program evaluation examinations in which the items are changed less frequently. Moreover, the items measure the cognitive skills that have great generality across the domain of the content, making it difficult and inefficient to drill for specific items. Teachers who do this rather than implement strong standards-based programs are under-serving the children. An example of this comes from the grade 4 English language arts test. A few administrations ago, we had a reading passage on Picasso. The students were asked questions regarding their abilities to understand the passage. After the administration, some teachers wrote to us that they had purchased materials on cubism. We, however, were not planning to have future passages on cubism, but rather passages that required some reading skill. Teaching in this way to the test is of no benefit to the students. Teaching to the learning standards from which each test form is derived is much more valuable.

Q 17: Does the use of passages from literature that may be familiar to some students provide an unfair advantage?

A: No. Two studies were performed in 1999 on the elementary English language arts test results to determine if students had gained an advantage on the test by prior exposure to one of the reading passages in instructional materials.

The first study analyzed scale scores for a sample of the total students tested. The differences between the whole scale score and a scale score without the passage that was familiar to some students were compared for exposed schools and schools matched by 1998 grade 3 reading passing rates and community type. No differences were found.

The second study predicted performance on the passage that was familiar to some students on the basis of performance on all other parts of the test and performance on the State's 1998 grade 3 reading test. Actual performance was subtracted from predicted performance. These differences, called residuals, were compared for students from exposed schools and for the matched unexposed schools. Again, no differences were found. Therefore, no unfair advantage was found to exist on the 1999 elementary English language arts test in relation to prior exposure to a reading passage.

Q 18: Why is a higher raw score required to achieve the same final score on some forms of the same Regents examinations, e.g., the June 2000 Regents Comprehensive Examination in English?

A: Every form of a new or revised Regents examination is equated through field-testing. This means that questions on each form of a test were given as a field test with other questions of known difficulty. When field test results indicate that the form is somewhat easier, it requires a higher raw score to demonstrate the same level of skill on this form as on a more difficult form.

Q 19: On the Regents Comprehensive Examination in English, what is the relative weighting of the objective questions and the essay questions?

A: According to the test design, the total essay score is given twice as much weight as the total multiple-choice score.

Q 20: What is component retesting, how does it work, and why do we need it?

A: Component retesting is part of an overall systemic program of prevention and intervention strategies for students who are at risk of not meeting the State learning standards. It is available for any senior student who has failed a Regents examination twice and has scored at least 48 on an examination. Schools can analyze the results of a student's tests and determine specific areas of the standards where he/she needs additional instruction.

In English, there are two component tests depending on areas of the standards for which students are deficient. The component tests are administered over five consecutive days – one class period per day. On day one and day two of one of

the English component tests, students write extended essays based on listening and/or reading passages. On the other three days, students answer multiple-choice questions and write shorter essays. For the other English component test, this schedule is reversed. The score on each component test is an accumulation of the five days of testing.

For mathematics, there are four possible components – each component is given on two consecutive days – one class period each day. Students answer both multiple-choice and solve multi-step problems.

The Department took specific steps to ensure that the total component tests are as rigorous as a section of the full Regents examination. The Educational Testing Service (ETS) was hired to ensure that the items on component tests were judged by New York State teachers to be scored at the appropriate level of difficulty and field tests were conducted to guarantee that the component retest questions were as difficult as a Regents examination.

In May 2001, schools across the State implemented component retesting in English and mathematics for the first time. Some press accounts have inaccurately characterized the component retests as "dumbing down" the Regents examinations. Our data does not support this conclusion. It is important to recognize and understand that component retesting provides students who are close to meeting Regents standards an additional opportunity to meet those standards.

Q 21: The State Assessment Panel has reviewed and recommended to the Commissioner alternative assessments to the Regents examinations. What has the Commissioner approved, and how often does the Panel meet to review proposed alternatives?

A: The Commissioner has approved a number of alternative assessments that can be taken in lieu of the Regents Comprehensive Examination in English and the Mathematics A Regents Examination. These approved alternatives are listed in Attachment C.

The State Assessment Panel usually meets twice a year (spring and fall) to review proposed alternatives. All approved alternatives must meet the following criteria:

- Be aligned with the New York State standards for that subject and be at least as rigorous as the corresponding required Regents examination;
- Meet technical criteria for validity, reliability and freedom from bias. (At a minimum, the assessment under consideration must: document relationship to domain or learning standards; document reliability and inter-rater reliability, as appropriate; have standard rubrics, as appropriate; document test development process; document procedure for establishing test performance standards, as appropriate; and document equating procedures or methods to insure comparability of forms.);

- Be externally developed and administered under secure conditions (i.e., the assessment cannot be developed exclusively by the teachers in the school nor can they have previous knowledge of the specific examination questions); and
- Be available for use by any school or school district in New York State.

Q 22: Will we be able to see all future questions on the grades 4 and 8 tests?

A: Yes, beginning with the 2002 tests.

Q 23: When can we expect to get the results from the grades 4 and 8 tests?

A: Results will be returned to schools within a few weeks of the testing dates.

Attachment A

The New York State Testing Program

DRAFT SCHEDULE FOR ELEMENTARY AND INTERMEDIATE STATE ASSESSMENTS FOR 2002

Exam	Administration Dates	Make-up Dates
Grade 4 English Language Arts	Monday, January 28-Friday, February 1*	Monday, February 4-Wednesday, February 6
Grade 4 Mathematics	Tuesday, May 7–Thursday, May 9	Friday, May 10–Tuesday, May 14
Grade 4 Elementary-Level Science	Any time in May	Any time in May
Grade 5 Elementary-Level Social Studies	Wednesday, November 14–Thursday, November 15	Friday, November 16–Monday, November 19
Grade 8 English Language Arts	Monday, March 4-Friday, March 8*	Monday, March 11-Tuesday, March 12
Grade 8 Mathematics	Tuesday, May 7–Wednesday, May 8	Thursday, May 9–Friday, May 10
Grade 8 Intermediate-Level Science Performance Test	Any time in May	Any time in May
Grade 8 Intermediate-Level Science Written Test	Any time between Wednesday, June 5 and Thursday, June 20	Any time between Thursday, June 6 and Thursday, June 20
Grade 8 Intermediate-Level Social Studies	Any time between Wednesday, June 5 and Thursday, June 20	Any time between Thursday, June 6 and Thursday, June 20
Intermediate-Level Technology Education	Any time between Wednesday, June 5 and Thursday, June 20	Any time between Thursday, June 6 and Thursday, June 20

*Public school districts and nonpublic schools should select specific dates (three for Grade 4, two for Grade 8) within the test administration windows specified above.

REGENTS EXAMINATION DATES FOR JANUARY AND JUNE 2002

Final dates for the January and June 2002 Regents examination periods have been established as follows:

Tuesday, January 22 – Friday, January 25, 2002

Tuesday, June 18 – Tuesday, June 25, 2002
(June 25, 2002 will be the rating day)

Attachment B

TEST DEVELOPMENT PROCESS

It takes approximately two years to develop a State assessment. The 19-step test development process used by the New York State Education Department has undergone thorough review and refinement in order to ensure that the assessments that are created are free from bias and are valid and reliable measures of student performance in relation to meeting the State learning standards. A quick review of the steps in the test development process follows:

TEST DEVELOPMENT STEPS	
1. Review syllabi/standards	11. Perform item analysis - estimate difficulty/discriminability - estimate distributions - estimate reliability - evaluate inter-rater reliability - review for sensitivity
2. Draw up test specifications	12. Review items and data
3. Solicit item writers	13. Field test forms - design sample - compose forms - solicit participants - print forms - distribute forms - review answer folders, test booklets - safeguard security - embed new items
4. Train item writers	14. Score objective field test questions
5. Publish prototypes of items/generic rubrics - publish sample items - seek district reviews	15. Read and score field test performance items - choose rangefinders - train readers - rate responses
6. Write items	16. Perform item analysis and test analysis
7. Review procedures and items - train committees - advise on policy/logistics - review content - advise on special populations	17. Submit to Statewide Review Committee
8. Pretest items - design sample - compose forms - solicit participants - print forms - distribute forms - review answer folders, test booklets - safeguard security	18. Conduct Standard Setting - determining the score point at which the State learning standards have been demonstrated
9. Score objective pretest questions	19. Develop new items (begin at Step 6)
10. Read and score pretest performance items - choose rangefinders - train readers - rate responses	

BEST COPY AVAILABLE

Attachment C

APPROVED ALTERNATIVE ASSESSMENTS

SUBJECT	APPROVED ALTERNATIVE ASSESSMENT	MINIMUM ACCEPTABLE SCORE	CORRESPONDING REGENTS EXAMINATION
ENGLISH	Advanced Placement Language and Composition Examination	3	REGENTS COMPREHENSIVE EXAMINATION IN ENGLISH
	Advanced Placement Literature and Composition Examination	3	
	Advanced International Certificate of Education (AICE) English Examination	E	
	International Baccalaureate English A1 Higher Level Examination	3	
	International Baccalaureate English A1 Standard Level Examination	4	
MATHEMATICS	Advanced International Certificate of Education (AICE) Mathematics Examination	*E	MATHEMATICS A OR SEQUENTIAL MATHEMATICS, COURSES I AND II
	Advanced Placement Calculus AB	*3	
	Advanced Placement Calculus BC Examination	*3	
	International Baccalaureate Mathematics Higher Level Examination	*3	
	International Baccalaureate Mathematics Methods Standard Level Examination	*4	
	International Baccalaureate Mathematics Studies Standard Level Examination	*4	
	SAT II Level IC	*470/**490	
	SAT II Level IIC Examination	*510/**550	
	International General Certificate of Secondary Education (IGCSE) Mathematics Examination	*A	

*Achieving the minimum acceptable score on any one of these mathematics examinations may be accepted as equivalent to passing with a 65 the Mathematics A Regents Examination or the Sequential Mathematics, Courses I and II Regents Examinations.

**Achieving this score on this mathematics examination may be accepted as equivalent to passing with a 65 the Sequential Mathematics, Courses I, II and III and Regents examinations.

Attachment D

GLOSSARY

Two sources are cited for the definitions included in this Glossary: *Standards for Educational and Psychological Testing, (copyright 1999* by the American Educational Research Association), reprinted with permission of the publisher. Definitions cited from this source will have a (1) after the term. The second source is *The Use of Tests as Part of High Stakes Decision Making for Students: A Resource Guide for Educators and Policymakers*, 2001, U.S. Department of Education's Office of Civil Rights. Definitions cited from this source will have a (2) after the term.

accommodation (2) – A change in how a test is presented, in how a test is administered, or in how the test taker is allowed to respond. This term generally refers to changes that do not substantially alter what the test measures. The proper use of accommodations does not substantially change academic level or performance criteria. Appropriate accommodations are made in order to level the playing field, i.e., to provide equal opportunity to demonstrate knowledge.

achievement levels/proficiency levels (2) – Descriptions of a test taker's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum, often labeled from "basic" to "advanced," that constitute broad ranges for classifying performance. See ***cut score***.

adjusted validity/reliability coefficient (1) – A validity or reliability coefficient—most often, a product-moment correlation—that has been adjusted to offset the effects of differences in score variability, or the unreliability of test and/or criterion. See ***restriction of range or variability.***

alternate forms (1) – Two or more versions of a test that are considered interchangeable, in that they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. ***Alternate forms*** is a generic term used to refer to any of three categories. ***Parallel forms*** have equal raw score means, equal standard deviations, equal error structures, and equal correlations with other measures for any given population. ***Equivalent forms*** do not have the statistical similarity of parallel forms, but the dissimilarities in raw score statistics are compensated for in the conversions to derived scores or in form-specific norm tables. ***Comparable forms*** are highly similar in content, but the degree of statistical similarity has not been demonstrated. See ***linkage.***

anchor test (1) – A common set of items administered with each of two or more different forms of a test for the purpose of equating the scores obtained on these forms.

assessment (1) – Any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs.

bias (2) – In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct irrelevant components of test scores. Bias usually favors one group of test takers over another.

calibration (1) – **1.** In linking test score scales, the process of setting the test score scale, including mean, standard deviation, and possibly shape of score distribution, so that scores on a scale have the same relative meaning as scores on a related scale. **2.** In item response theory, the process of determining the parameters of the response function for an item.

construct (1) – The concept or the characteristic that a test is designed to measure.

construct irrelevance (1) – The extent to which test scores are influenced by factors that are irrelevant to the construct that the test is intended to measure. Such extraneous factors distort the meaning of test scores from what is implied in the proposed interpretation.

construct underrepresentation (1) – The extent to which a test fails to capture important aspects of the construct that the test is intended to measure. In this situation, the meaning of test scores is narrower than the proposed interpretation implies.

construct validity (1) – The term used to indicate that the test scores are to be interpreted as indicating the test taker's standing on the psychological construct measured by the test. A construct is a theoretical variable inferred from multiple types of evidence, which might include the interrelations of the test scores with other variables, internal test structure, observations of response processes, as well as the content of the test. In the current standards, all test scores are viewed as measures of some construct, so the phrase is redundant with validity. The validity argument establishes the construct validity of a test. See ***construct, validity argument.***

content validity (2) – Validity evidence which analyzes the relationship between a test's content and the construct it is intended to measure. Evidence based on test content includes logical and empirical analyses of the relevance and representativeness of the test content to the defined domain of the test content to the defined domain of the test and the proposed interpretations of test scores.

criterion-referenced test (1) – A test that allows its users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others. Examples of criterion-referenced interpretations include comparison to cut scores, interpretations based on expectancy tables, and domain-referenced score interpretations.

cut score (1) – A specified point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point. See ***performance standard.***

derived score (1) – A score to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

equated forms (1) – Two or more test forms constructed to cover the same explicit content, to conform to the same statistical specifications, and to be administered under identical procedures (*alternate forms);* through statistical adjustments, the scores on the alternate forms share a common scale.

equating (1) – Putting two or more essentially parallel tests on a common scale. See ***alternate forms.***

fairness (1) – In testing, the principle that every test taker should be assessed in an equitable way.

field test (1) – A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting. A field test is generally more extensive than a pilot test. See ***pilot test.***

generalizability theory (1) – An extension of classical reliability theory and methodology in which the magnitudes of error from specified sources are estimated through the use of one or another experimental design, and the application of the statistical techniques of the analysis of variance. The analysis indicates the generalizability of scores beyond the specific sample of items, persons, and observational conditions that were studied.

inter-rater agreement (1) – The consistency with which two or more judges rate the work or performance of test takers; sometimes referred to as *inter-rater reliability*.

item (1) – A statement, question, exercise, or task on a test for which the test taker is to select or construct a response, or perform a task. See ***item prompt.***

item pool (1) – The aggregate of items from which a test or test scale's items are selected during test development, or the total set of items from which a particular test is selected for a test taker during adaptive testing.

item prompt (1) – The question, stimulus, or instructions that direct the efforts of examinees in formulating their responses to a constructed-response exercise.

item response theory (IRT) (1) – A mathematical model of the relationship between performance on a test item and the test taker's level of performance on a scale of the ability, trait, or proficiency being measured, usually denoted as θ. In the case of items scored 0 / 1 (incorrect/correct response) the model describes the relationship between θ and the item mean score (P) for test takers at level θ, over the range of permissible values of θ. In most applications, the mathematical function relating P to θ is assumed to be a logistic function that closely resembles the cumulative normal distribution.

linkage (1) – The result of placing two or more tests on the same scale, so that scores can be used interchangeably. Several linking methods are used: See ***equating, calibration, moderation, projection, and alternate forms.***

moderation (1) – In test linking, the term moderation used without a modifier, usually signifies statistical moderation, which is the adjustment of the score scale of one test, usually by setting the mean and standard deviation of one set of test scores to be equal to the mean and standard deviation of another distribution of test scores.

performance assessment (1) – Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied.

performance standard (1) – **1.** An objective definition of a certain level of performance in some domain in terms of a cut score or a range of scores on the score scale of a test measuring proficiency in that domain. **2.** A statement or description of a set of operational tasks exemplifying a level of performance associated with a more general content standard; the statement may be used to guide judgments about the location of a cut score on a score scale. The term often implies a desired level of performance. See ***cut score.***

pilot test (1) – A test administered to a sample of test takers to try out some aspects of the test or test items, such as instructions, time limits, item response formats, or item response options. See ***field test.***

predictive bias (1) – The systematic under- or over-prediction of criterion performance for people belonging to groups differentiated by characteristics not relevant to criterion performance.

projection (1) – In test scaling, a method of linking in which scores on one test (X) are used to predict scores on another test (Y). The projected Y score is the average Y score for all persons with a given X score. Like regression, the projection of test Y onto test X is different from the projection of test X onto test Y. See ***linkage.***

raw score (1) – The unadjusted score on a test, often determined by counting the number of correct answers, but more generally a sum or other combination of item scores. In item response theory, the estimate of test taker proficiency, usually symbolized θ, is analogous to a raw score although, unlike a raw score, its scaling is not arbitrary.

reliability (1) – The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors or measurement for a given group. See ***generalizability theory.***

restriction of range or variability (1) – Reduction in the observed score variance of an examinee sample, compared to the variance of the entire examinee population, as a consequence of constraints on the process of sampling examinees.

scale (1) – **1.** The system of numbers, and their units, by which a value is reported on some dimension of measurement. Length can be reported in the English system of feet and inches or in the metric system of meters and centimeters. **2.** In testing, *scale* sometimes refers to the set of items or subtests used in the measurement and is distinguished from a test in the type of characteristic being measured. One speaks of a test of verbal ability, but a scale of extroversion-introversion.

scale score (1) – See ***derived score.***

scaling (1) – The process of creating a scale or a scale score. Scaling may enhance test score interpretation by placing scores from different tests or test forms onto a common scale or by producing scale scores designed to support criterion-referenced or norm-referenced score interpretations. See ***scale.***

scoring rubric (1) – The established criteria, including rules, principles, and illustrations, used in scoring responses to individual items and clusters of items. The term usually refers to the scoring procedures for assessment tasks that do not provide enumerated responses from which test takers make a choice. Scoring rubrics vary in the degree of judgment entailed, in the number of distinct score levels defined, in the latitude given scorers for assigning intermediate or fractional score values, and in other ways.

speededness (1) – A test characteristic, dictated by the test's time limits, that results in a test taker's score being dependent on the rate at which work is performed as well as the correctness of the responses. The term is not used to describe tests of speed. Speededness is often an undesirable characteristic.

standards-based assessment (1) – Assessments intended to represent systematically described content and performance standards.

technical manual (1) – A publication prepared by test authors and publishers to provide technical and psychometric information on a test.

test (1) – An evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process.

test development (1) – The process through which a test is planned, constructed, evaluated, and modified, including consideration of content, format, administration, scoring, item properties, scaling, and technical quality for its intended purpose.

test documents (1) – Publications such as test manuals, technical manuals, user's guides, specimen sets, and directions for test administrators and scorers that provide information for evaluating the appropriateness and technical adequacy of a test for its intended purpose.

test manual (1) – A publication prepared by test developers and publishers to provide information on test administration, scoring, and interpretation and to provide technical data on test characteristics. See ***user's guide.***

test specifications (1) – A detailed description for a test, often called a test blueprint, that specifies the number or proportion of items that assess each content and process/skill area; the format of items, responses, and scoring rubrics and procedures; and the desired psychometric properties of the items and test such as the distribution of item difficulty and discrimination indices.

user's guide (1) – A publication prepared by the test authors and publishers to provide information on a test's purpose, appropriate uses, proper administration, scoring procedures, normative data, interpretation of results, and case studies. See ***test manual.***

validity (1) – The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test.

validity argument (1) – An explicit scientific justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores.

weighted scoring (1) – A method of scoring a test in which the number of points awarded for a correct (or diagnostically relevant) response is not the same for all items in the test. In some cases, the scoring formula awards more points for one response to an item than for another.